Homework Assignment #4
Due at midnight Friday 12/2

Part-1

For this homework, you are to write a program that implements both naive Bayes and TAN (tree-augmented naive Bayes).
main file: *bayes.py*
script: *bayes*

Part-2

For this part, use stratified 10-fold cross validation on the *chess-KingRookVKingPawn.arff* data set to compare naive Bayes and TAN. Be sure to use the same partitioning of the data set for both algorithms. Report the accuracy the models achieve for each fold and then use a paired t-test to determine the statistical significance of the difference in accuracy. Report both the value of the t-statistic and the resulting p value.

Code included in file: *bayes_stratified_cross_validation.py*

Accuracy of models on each fold:

Accuracy	**TAN** (Group 1)	**Naive Bayes** (Group 2)
Fold-1 Accuracy	0.8906250000000000	0.7906250000000000
Fold-2 Accuracy	0.8656250000000000	0.7875000000000000
Fold-3 Accuracy	0.8843750000000000	0.7875000000000000
Fold-4 Accuracy	0.9218750000000000	0.8156250000000000
Fold-5 Accuracy	0.9062500000000000	0.8218750000000000
Fold-6 Accuracy	0.9156250000000000	0.8187500000000000
Fold-7 Accuracy	0.8906250000000000	0.8406250000000000
Fold-8 Accuracy	0.9404388714733543	0.8369905956112853
Fold-9 Accuracy	0.9874608150470220	0.9310344827586207
Fold-10 Accuracy	0.9937106918238994	0.9811320754716981

Paired T-test Result:

Mean of differences is 0.078495
Variance of differences is 0.000912
Standard deviation of differences is 0.030203
t value is 8.218621
Degree of freedom is 9
p value is 0.000018
The 95% confidence interval of the differences is from 0.056890 to 0.100101
By conventional criteria, this difference in accuracy is considered to be extremely statistically significant.

Data Summary:

There are 10 samples for both groups
Group 1 has

- mean 0.919661
- variance 0.001840 and
- standard deviation 0.042901

Group 2 has

- mean 0.841166
- variance 0.004170 and
- standard deviation 0.064572